Better together: modern methods plus traditional thinking in NP alignment

Ádám Kovács[1,2], Judit Ács[1,2], András Kornai[2], Gábor Recski[1,3]
[1]BME Dept. of Automation and Applied Informatics, [2]SZTAKI Institute of Computer Science, [3]Apollo.AI
lastname.firstname@aut.bme.hu, andras@kornai.com, gabor@apollo.ai

Abstract

We study a typical intermediary task to Machine Translation, the alignment of NPs in the bitext. After arguing that the task remains relevant even in an end-to-end paradigm, we present simple, dictionary- and word vector-based baselines and a BERT-based system. Our results make clear that even state of the art systems relying on the best end-to-end methods can be improved by bringing in old-fashioned methods such as stopword removal, lemmatization, and dictionaries

Keywords: NP-alignment, rule-based, BERT, hybrid

1. Introduction

As the state of the art in machine translation (MT) is now dominated by end-to-end (e2e) neural architectures, the task of aligning phrases of parallel corpora has been neglected. In fact, the tides have turned to such an extent that discovery of intermediate structure, hitherto the standard approach, now requires special justification. Since this is obviously not the place to survey the entire state of the art, we confine ourselves to a few remarks, beginning with the obvious: e2e, at least as currently practiced, is not structure free.

In MT, e2e systems are built on two pivotal structures: the segmentation of the speech stream into *sentences*, and the segmentation of the sentences into *words*. As we shall see, the arguments in favor of these structural levels apply, with the same force, to the intermediary level that is our chief concern here, *phrases*. One simply cannot train MT systems without sentence-aligned corpora (our point (1) below), and the embeddings that both e2e and more modular systems rely on assuming subsentential units such as words, word-parts, or characters (our (2) below). Training a truly e2e system on naturally occurring speech streams (seq2seq transformation of acoustic frames) is not on the horizon.

While somewhat obscured by the fact that the segmentation effort is hidden in the preprocessing, there is no denying that e2e, as practiced today and in the foreseeable future, fundamentally relies on conventionally built structures that already pre-package a great deal of the structural information linguists take to be relevant. To be sure, characters are not exactly phonemes, and word pieces are not exactly morphemes, but the expectation is that the better these systems become the more the units (which in the e2e paradigm have to be teased out of the system by specialized probes) will resemble phonemes, morphemes, etc.

In this paper we argue that using traditional, structure-based thinking, even about tasks such as MT that seamlessly fit the e2e paradigm, can lead to improvements in the state of the art. In the rest of this Introduction, we enumerate what we see as the main reasons for positing intermediate structures. We describe our data in Section 2.; the traditional and contemporary methods in Section 3.; and in Section 4. we present the results that justify our conclusion that hybrids using both traditional and modern components improve performance.

1. *Structure facilitates data gathering.* Since the data generally follows Zipf's Law, there is a non-negligible heavy tail that requires one-shot or even zero-shot learning. This is especially clear in MT, where data must be aggregated in word or subword units for training, but there are many other problems, even MT for low-resource languages, where we simply don't have sufficient data for taking full advantage of e2e capabilities.

2. *Structures subdivide the task.* For any sequence labeling task that transform some sequence r_i of input units to some sequence t_k of output units, to the extent some intermediate layer of representation s_j can be established, this subdivides the task into two transformations that are both individually better learnable than their convolution. Further, in cases where the s_j are linguistic signs, linking sound and meaning in an arbitrary fashion, no gradients can reasonably be expected to flow through the s_j.

3. *Structures are multifunctional pivots.* Those structures that are treated as standard in linguistics generally have relevance for several domains, not just sound and meaning. For example, the state-of-the-art (SOTA) object recognizer, YOLO9000 (Redmon et al., 2016), structures visual images by words and phrases, and in the cognitive science literature (Rosch, 1975; Lakoff, 1987; Gärdenfors, 2000) it has long been argued that meanings extend to other sensory domains (vision in particular) by means of prototypes.

   In tasks like captioning (Karpathy et al., 2014) reasons (1-3) appear together: the tail is so heavy that for the most part even one-shot learning is out of the question, gradients don't flow through, and practically all results have to be assembled compositionally from intermediate structures.

4. *Structures facilitate explanation.* By their very nature, e2e systems are black boxes, but there are many situations where ethical and practical considerations demand a human-understandable explanation why the system made a particular choice. Generally, tracing through different levels of representation goes a long way toward the eXplainable AI (XAI) goal.

5. *Structure facilitates debugging* Closely linked to our previous point, e2e systems, especially those driving the state of the art, typically embody the work of many people and often millions of GPU hours. In other fields of engineering we rarely encounter major systems that do not include some sort of inspection hatches and instrumentation for human observers, and it goes against the grain of centuries of engineering experience to assume that here is a situation where we could do without.

Let us briefly consider how the points made here apply to phrases. 1. Phrases follow Zipf's law. 2. Identified phrases clearly subdivide the translation task. This is especially transparent for named entities (NEs) which, as fixed units, generally do not require translation. For example the English LOC *San Diego* is translated to Hungarian as *San Diego* rather than *Szent Jakab*, which would be the correct translation for PER (but the person is called *Saint James the Great* in English, where *San Diego* or *Santiago* is restricted to LOCs).

As for 3, phrases are units in phonology both in terms of driving intonation contours and in terms of pauses/breath groups; in grammar, where verbal complements like subjects and objects are expressed in terms of phrases; in syntax, where rigid phrase-internal word order is often seen even in otherwise "free word order" languages; and in semantics, where phrases often designate entities in a non-compositional fashion, e.g. *Action Française* doesn't refer to some kind of French action but rather to an ultra-right political movement. In particular, integration with knowledge-based systems is virtually impossible without NE keys.

Finally, for 4-5, in the course of this work we have noticed that cases of phrase mismatch are predictive of Google Translate errors. Our corpus is Orwell's *1984* (see Section 2.), which contains sentences like *It was part of the economy drive in preparation for Hate Week*, with gold translation *Ez is része volt a takarékossági versenynek, amellyel a Gyűlölet Hetére készültek*.

On the whole, Google Translate does a commendable job of translating the Hungarian back to English. But when our aligners (see Section 3.) fail, we are more likely to see a translation failure, something that affects the meaning, not just the style. Here we get *This was also part of the austerity competition that was being prepared for Hate Week.* In the original, and in the Hungarian gold translation, we have the subject *It* (translated back as *This* – we don't consider this an error), the goal *Hate Week*, and the object *economy drive* (translated back as *austerity competition* – again we don't consider this an error). However, in the original the subject is in preparation of the goal, whereas in the Google translation it is the entire object that is in preparation. The distinction is subtle, but errors of this kind generally escalate to a full failure in semantic tasks like recognizing textual entailment (RTE) (Dagan et al., 2006), see also `https://aclweb.org/aclwiki/Textual_Entailment_References`.

2. Data

Our core data set, a manually translated and word-aligned corpus of Orwell's *1984*, was created as part of the MULTEX-East project (Erjavec, 2004). A phrase-level alignment between English and Hungarian noun phrases (NPs) (Recski et al., 2010) was presented in [1]. This dataset contains 6567 sentence pairs, or bi-sentences, with 25,561 English and 22,408 Hungarian NPs. Only NPs that are not contained by a higher level NP (i.e. top-level NPs) are annotated. By today's standards, the dataset is tiny, but as we noted in (Recski et al., 2010), "NP alignment is a challenging problem, capable of rapidly exposing flaws both in the word-alignment and in the NP chunking algorithms one may bring to bear". There, the same dataset was used to train and test a GIZA-style aligner (Och and Ney, 2003) which carried most of the workload, while here the bulk of the work is carried by the independently trained MUSE and BERT, with the corpus used only for adaptation. But for tasks involving low resource languages, and languages with a great deal of morphology (the two often come hand in hand, though we consider Hungarian to be medium-, rather than low-resourced) it is not just stemming, but also the case marking that decides which NP fills which slot, remain relevant.

While the task of NP alignment usually involves aligning NPs in a bi-sentence and possibly based on wider context, in this paper we reduce it to the simpler task of deciding for a pair of NPs whether they should be aligned or not, based only on the NPs itself and knowledge of the fact that they are within the same bi-sentence. We therefore extract all alignment candidates from the 1984 corpus, i.e. all pairs of English and Hungarian NPs such that their sentences are translations of each other, along with a ground truth label indicating whether these NPs should in fact be aligned with each other. The entire dataset contains 121,783 NP pairs (or 18.5 per sentence) of which 18 789 (2.9 per sentence) are labeled as alignment pairs.

To experiment with a simple rule-based solution we used dictionaries from (Ács et al., 2014)(Ács et al., 2013)[2] and (Kornai and Tóth, 1997)[3] as well as the MUSE multilingual word embeddings (Conneau et al., 2017)(Conneau et al., 2017)[4] and by training a simple classifier based on BERT representations (Devlin et al., 2018).[5] All systems described in this paper, as well as all scripts necessary to reproduce our results are available under an MIT license at `https://github.com/adaamko/np_alignment`.

3. Methods

3.1. MUSE

Our simplest method evaluates the similarity of pairs of English and Hungarian noun phrases by mapping their words

[1] `https://hlt.bme.hu/en/resources/1984-corpus`

[2] `https://github.com/juditacs/wikt2dict`

[3] `https://hlt.bme.hu/en/resources/hokoto`

[4] `https://github.com/facebookresearch/MUSE`

[5] We use the PyTorch implementation `https://github.com/huggingface/transformers`

to the universal embedding space of MUSE vectors. We obtain bag-of-words representations of NPs by removing stopwords using NLTK(Bird et al., 2009) and lemmatize using spacy(Honnibal and Montani, 2017) for English and emmorph(Novák et al., 2016) for Hungarian. We leave unchanged NPs that contain only stopwords. Then, let (E, H) denote sets of MUSE vectors corresponding to the words of any pair of English and Hungarian NPs belonging to the same bi-sentence. We approximate the likelihood of aligning the two NPs by the maximum cosine similarity between any two words of the two NPs:

$$S(E,H) = \max_{(w_E,w_H)\in E\times H} \frac{w_E w_H}{|w_E||w_H|}$$

When aligning NPs of a sentence pair, we add edges between all pairs of NPs where the above similarity is above a given threshold. Based on the baseline's performance on the training dataset, we set this threshold to 0.46. Figure 1 shows precision, recall, and F-score values on the training set as a function of the threshold.

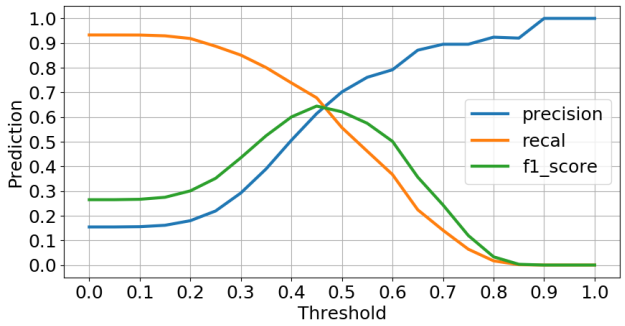

Figure 1: Performance of the MUSE baseline on the training at various thresholds

If all the words in an NP are outside the embedding's vocabulary (OOV), we add an edge iff there's at least one pair of English and Hungarian words whose Levenshtein distance is less than 4. This fallback slightly improves recall by matching pairs of proper nouns such as *Oceania* and *Óceánia*. We shall use the same fallback for OOVs in our dictionary-based method described in Section 2..

3.2. BERT

Our second method maps English and Hungarian NPs to vectors using the multilingual BERT language model. For each pair of NPs in a bi-sentence, we obtain a BERT sentence embedding by concatenating the two sentences (with a sentence boundary symbol in-between) before using them as input to the pretrained BERT model and extracting weights by summing up its last 4 hidden layers. Finally, we keep only vectors corresponding to words of the two NPs and feed them into an LSTM layer (Hochreiter and Schmidhuber, 1997), the output of which is then used by a linear layer that predicts the probability of the two NPs being aligned. As loss function we used negative log likelihood loss, the classifier was trained using the Adam optimizer, with a starting learning rate of 0.01 and early stopping to avoid overfitting. Since there are approximately 6 times more negative samples in the data than true edges, we experimented both with weighted loss functions and with over- and under-sampling. The best results were achieved by oversampling positive examples.

Method	Precision	Recall	F-score
always yes	15.43	100	26.73
surface	22.30	38.27	28.18
MUSE	63.51	66.29	64.87
MUSE+surface	63.52	67.96	65.66
BERT	67.06	**77.20**	71.77
Dict	77.49	72.01	74.65
Dict+surface	**78.08**	76.66	**77.36**

Table 1: Maximum precision, recall and F-score of the systems.

3.3. Dictionary-based alignment

Our dictionary-based system uses English-Hungarian translation pairs from the Wikt2dict and Hokoto dictionaries described in Section 2.. For each NP pair we first perform stopword filtering and lemmatization as described in 3.1., then retrieve the list of all Hungarian equivalents for all words of the English NP. Then, if there is any pair of words in the two NPs such that the Hungarian word is among the translations of the English word, we add an alignment edge between the NPs. For words that are at least 5 characters long, a Levenshtein distance not greater than 3 is enough for the words to be considered a match. Our initial experiments let us determine these parameters and that best results are achieved if even a single pair of corresponding words triggers adding an alignment edge between the two phrases. For words not in the dictionary (OOVs) we fall back to the surface-based baseline described in Section 3.1..

4. Results

We split the set of labeled NP pairs extracted from the 1984 dataset into train and test portions. The training dataset was used to train the BERT-based system as well as to find optimal parameters of the other two baselines. The test dataset contains 24,357 NP pairs, of which 3,758 (15.43%) are connected by a gold alignment edge. Table 1 shows the performance of each system introduced in Section 3.. In case of MUSE and the dictionary-based methods, we tested the systems with and without the surface-based baseline as a fallback for OOV words. We also evaluate this baseline on its own (*surface*) and the system that would align all NP pairs within a bi-sentence (*always yes*). The dictionary-based method with fallback yields the highest performance with an F-score of 77.36.

Next we experimented with various voting schemes, i.e simple rules to determine final labels based on the labels provided by some subset of our systems. Table 2 shows the results on the three combinations that outperformed individual systems on at least one figure of merit. Combining BERT and Dict+surface by AND and OR, i.e. adding edges between NPs iff at least one or iff both systems decided to do so, yield very high precision and recall values, respectively. This indicates that the two systems, while achieving comparable performance on their own, actually iden-

tified quite different subsets of edges; in fact, they assign the same label to only 88.2% of all edges. Also, this provides us with systems that can achieve over 90% precision or recall in scenarios where one is to be favored over the other. The best overall results are achieved by majority vote among the three systems, yielding an F-score of 80.51.

Method	Precision	Recall	F-score
BERT $\vee$ Dict+surface	62.61	90.77	74.10
BERT $\wedge$ Dict+surface	92.33	63.09	74.96
3-way vote	82.30	78.79	80.51

Table 2: Performance of hybrid systems

5. Conclusion

The results presented in Table 2 make clear that even SOTA systems relying on the best e2e methods can be improved by bringing in old-fashioned methods such as stopword removal, lemmatization, and dictionaries. Since the effectiveness of these methods has been known for decades (for stopwords, see (Luhn, 1959); for lemmatization see (Porter, 1980); and for dictionaries see (McNaught, 1988; Miller, 1995)) our results are hardly surprising, except for clearly going against the grain of the prevailing e2e philosophy.

The fact that such methods actually improve SOTA systems has already been observed (Lauscher et al., 2019; Zhang et al., 2019), and the value of intermediate representations has been eloquently argued (Bengio et al., 2013). But the main implication, somewhat unpleasant for e2e, that there is still a great deal of value in painstakingly built gold resources, has not been fully drawn.

We conclude with a simple prediction: on cross-modal tasks, whether the non-linguistic modality is logic, as in RTE; image, as in object labeling; audio, as in speech recognition; or video, as in sign language recognition; intermediate representations remain essential, since gradients will not cross the modality barrier. Since robotics involves many such tasks, NLP is better off by building incrementally on the results of earlier paradigms.

6. Acknowledgments

Work partially supported by the ÚNKP-19-3 New National Excellence Program of the Ministry for Innovation and Technology; by 2018-1.2.1-NKP-00008: Exploring the Mathematical Foundations of Artificial Intelligence; by Project no. FIEK_16-1-2016-0007 (implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the Centre for Higher Education and Industrial Cooperation - Research infrastructure development (FIEK_16) funding scheme) and by National Research, Development and Innovation Office grant NKFIH #120145 'Deep Learning of Morphological Structure'. We thank Ken Church (Baidu) for the inspiration to step out of the e2e mainstream and for many helpful comments.

7. Bibliographical References

Ács, J., Pajkossy, K., and Kornai, A. (2013). Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria. Association for Computational Linguistics.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Trans. PAMI*, 35(8):1798–1828.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Dagan, I., Glickman, O., and Magnini, B. (2006). The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, volume 3944 of *LNCS*, pages 177–190. Springer.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Erjavec, T. (2004). MULTEXT-east version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In Maria Teresa Lino, et al., editors, *Fourth Inernational Conference on Language Resources and Evaluation (LREC 2004)*, pages 1535–1538, Paris. European Language Resources Association (ELRA).

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. MIT Press.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780, 11.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Karpathy, A., Joulin, A., and Li, F. F. (2014). Deep fragment embeddings for bidirectional image sentence mapping. In Z. Ghahramani, et al., editors, *Advances in Neural Information Processing Systems 27*, pages 1889–1897. Curran Associates, Inc.

Kornai, A. and Tóth, G. (1997). Gépi ékezés. *Magyar Tudomány*, 42(4):400–410.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. University of Chicago Press.

Lauscher, A., Vulič, I., Ponti, E. M., Korhonen, A., and Glavaš, G. (2019). Informing unsupervised pretraining with external linguistic knowledge. *arXiv preprint arXiv:1909.02339*.

Luhn, H. P. (1959). Keyword-in-Context Index for Technical Literature (KWIC Index). *American Documentation*, 11(4):288–295.

McNaught, J. (1988). Computational lexicography and computational linguistics. *Lexicographica*, 4:19–33.

Miller, G. A. (1995). Wordnet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Novák, A., Siklósi, B., and Oravecz, C. (2016). A new integrated open-source morphological analyzer for Hun-

garian. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Recski, G., Rung, A., Zséder, A., and Kornai, A. (2010). Np alignment in bilingual corpora. In Nicoletta Calzolari, et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 3379–3382, Valletta, Malta, may. European Language Resources Association (ELRA).

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology*, 104(3):192–233.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., and Liu, Q. (2019). ERNIE: Enhanced Language Representation with Informative Entities. *arXiv preprint arXiv:1905.07129*.

8. Language Resource References

Ács, Judit and Pajkossy, Katalin and Kornai, András. (2014). *Wikt2dict multilingual dictionaries*. SZTAKI HLT.

Conneau, Alexis and Lample, Guillaume and Ranzato, Marc'Aurelio and Denoyer, Ludovic and Jégou, Hervé. (2017). *MUSE multlilingual dictionaries*.

Kornai, András and Tóth, Gábor. (1997). *HoKoTo English-Hungarian dictionary*.

Gábor Recski and András Rung and Attila Zséder and András Kornai. (2010). *1984 NP-aligned corpus of English and Hungarian*.